Developing an Efficient Topic Modeling Framework for Image Classification

SHORT SYNOPSIS

Submitted in partial fulfillment of the requirement of the degree of

DOCTOR OF PHILOSOPHY

to

Manav Rachna International Institute of Research and Studies

(Deemed to be University)

by

Ms. Astha Goyal
Registration number: 21/Ph.D/046

Under the supervision of

Supervisor
Dr. Indu Kashyap
Professor, CSE

SCHOOL OF ENGINEERING & TECHNOLOGY

MANAV RACHNA INTERNATIONAL INSTITUTE OF RESEARCH & STUDIES

SECTOR-43, SURAJKUND –DELHI ROAD, FARIDABAD – 121001

SCHOOL OF ENGINEERING & TECHNOLOGY

Certificate

I hereby declare that the short synopsis "Developing an Efficient Topic Modeling Framework for Image Classification" being submitted in partial fulfillment of the requirement for the degree of Doctor of Philosophy in Computer Science under the School of Engineering of Manav Rachna International Institute of Research & Studies, Faridabad during the academic year 2023-24.

It is a bona fide record of original work carried out under the supervision of Dr. Indu Kashyap. It has not been submitted in full or part to any other university or institute for a degree or diploma. Further, it is certified that the short synopsis is plagiarism-free, and the proof has been submitted.

Astha Goyal 23/11/23

(Ms. Astha Goyal)

Registration number: 21/Ph.D/046

23/11/23

Signature of the Supervisor

Dr. Indu Kashyap

Professor, CSE

School of Engineering & Technology

23/11/23

(Chairperson, DRC)

23/11/23

(Director, CDPP)

ABSTRACT

Topic modeling is a powerful tool with many applications in fields such as text summarization, sentiment analysis, and reviewing of vast literature. In such contexts, topic models are used to discern and extract the relevant topics or themes within a given corpus of text, enabling the creation of summaries, the analysis of sentiment, or the synthesis of research findings. However, topic models are not limited to application over textual data. They can also be employed over other forms of data, such as multimedia content, for tasks such as classification and captioning.

This research investigates the effectiveness of utilizing topic models on image captions to classify images. Existing approaches for image classification focus solely on visual information and fail to use information from other modalities, as well as abstract information hinted by latent topics corresponding to textual descriptions/captions for an image. This study explores the utilization of textual modality associated with images, using topic models for image classification. It reviews topic modeling, image captioning, and classification methods. A framework is proposed, leveraging 'K' topics and image captions. The approach aims to enhance image classification by extracting nuanced topic information, benefiting image recognition, search, and retrieval capabilities.

The findings of this research have the potential to showcase topic modeling as a robust tool for image classification. The proposed framework will be evaluated at different levels using evaluation metrics. A combination of internal and external quality metrics will be used for topic modeling, including Topic Classification, Topic Significance, Topic Coherence, Topic Diversity, Adjusted Rand Index, and the Adjusted Mutual Index. BLEU, ROUGE, METEOR, and CIDER will be used for image captioning. Whereas Confusion Matrix, Accuracy, Precision, Recall, and F1 score will be employed for evaluating image classification results.

Keywords: Hyperparameter Tuning, Image Classification, Latent Dirichlet Allocation, Multimodal Approach, Topic Modeling

Contents

1. Introduction 1
1.1. Topic Modeling 1
1.2. Topic Models 2
1.3. Evaluation of Topic Models 3
1.4. Hyperparameter Tuning 4
1.5. Topic Modeling for Images 4
2. Literature Review 5
2.1. Topic Models 5
2.2. Hyperparameter Tuning: Optimal Number of Topics 'K' 12
2.3. Image Captioning using Topic Modeling 17
2.4. Image Classification using Topic Modeling 21
2.5. Research Gaps 23
3. Problem Definition 24
4. Objectives of Research 24
5. Research Methodology 24
6. Expected outcome of the research 29
References 30
Appendix 35

1. Introduction

In Computer Vision (CV), image classification is a task where a machine learning algorithm is trained to identify and classify images into predefined categories. In this task, the algorithm is trained over a large dataset of labeled images, by which it learns to identify descriptor patterns and features that differentiate images across different classes.

Captioning an image is a related task in the CV that involves generating a textual description of an image. In this task, the algorithm is expected to synthesize a natural language description of the contents of an image. This task typically combines CV and Natural Language Processing (NLP) techniques to generate an accurate and descriptive caption.

Topic modeling is an NLP technique used to discover the underlying topics or themes within a collection of text documents. While it is not directly applicable to image classification and captioning, it can be used to improve performance in conjunction with these tasks. One approach to using topic modeling for image classification is identifying the key concepts or themes most relevant to a given set of images. This can be done by analyzing the text associated with the images, such as captions or tags, and identifying the most common topics or themes. Once these topics have been identified, they can inform the image classification algorithm as an additional feature or group similar images.

Although topic modeling alone may not be a direct solution for image classification, it can be employed as a supplementary approach to enhance its effectiveness. By discerning the most pertinent topics or themes associated with images, topic modeling can enhance the accuracy and pertinence of image classification tasks.

1.1. Topic Modeling

Topic modeling is the statistical process of identifying, classifying, and summarizing the theme of a collection of possibly unstructured documents in a corpus. Algorithms for topic modeling referred to as topic models, aim to unravel a concise descriptor or topic that provides insight into the document's theme. Information about topics can be utilized to search

relevant documents in a large corpus and effectively organize and cluster documents according to similarity, text summarization, and collaborative filtering. Topic models may also be employed over non-textual data, such as data from the study in genetic-images & social networks. For example, topic models can provide insight into demographic behavior on a social media platform by analyzing themes from posts. Observing these themes over time can provide insight into the areas of interest of people and the evolutionary trends of these interests (Bide, et al., 2021). “Topic models” are the algorithms used in topic modeling. Every topic model employs the “bag of words” methodology, which ignores the syntax and word order in favor of the frequency and variety of words. Using clustering algorithms, words are divided into groups according to how frequently they appear in the documents. After identifying topics that use these word groups, it is simple to categorize the documents.

1.2. Topic Models

Several versions of topic models have evolved. During the early 1990s, the inception of topic models was marked by the introduction of Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Analysis (PLSA), and Non-negative Matrix Factorization (NMF). These initial models laid the foundation for subsequent research endeavors. As the exploration of topic models progressed, the researchers advanced their investigations, culminating in developing an exemplary algorithm known as Latent Dirichlet Allocation (LDA), which stands as a state-of-the-art approach in this field. Depending on the algorithm's nature and the inference method, these can be broadly classified into probabilistic and non-probabilistic (Kherwa & Bansal, 2019). Out of these, LDA is the most popularly employed topic modeling algorithm. Being a generative algorithm, it views a document as a distribution over latent topics and formulates a topic as a distribution over words. These distributions reveal the underlying theme of the collection of documents. To identify the topic, LDA optimizes the Dirichlet priors ‘α’ and ‘β’ over the multinomial distributions of the topics and the words in the document iteratively (Cao, et al., 2009). The LDA model discovers the topic by determining the word-level correlation but cannot capture cross-topic correlation. In real-world topics, correlations are common, ignoring which limits the scalability of LDA to real-world data. It is essential to model topic correlations to achieve better topic discovery. Over a period, several variants of LDA have evolved, as shown in Figure 1.

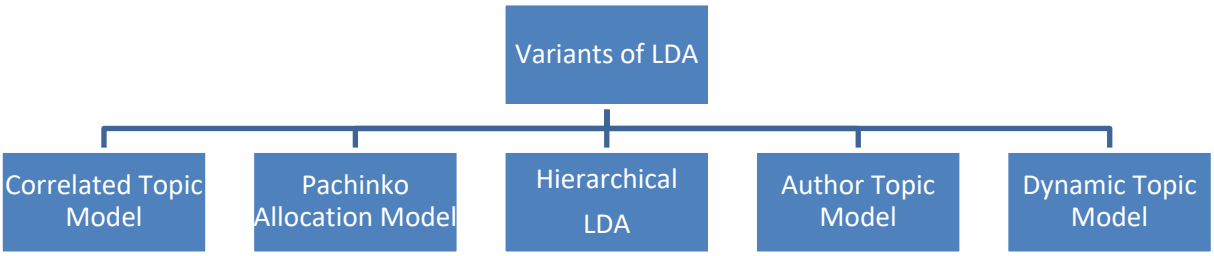

Figure 1: Variants of LDA (Source: Self)

1.3.Evaluation of Topic Models

The quality of topics generated by topic models can be assessed using evaluation metrics to quantify various topics, such as correlation, similarity, saliency, and relevance (Terragni, et al., 2021). A metric can serve as the objective for a Bayesian Optimization strategy (used for achieving convergence during training over corpora or for hyperparameter tuning) or for monitoring the performance of topic models. In contrast, the model gets optimized over a differing objective. Across the literature, based on the aspects investigated, evaluation metrics have been broadly classified, as shown in Figure 2.

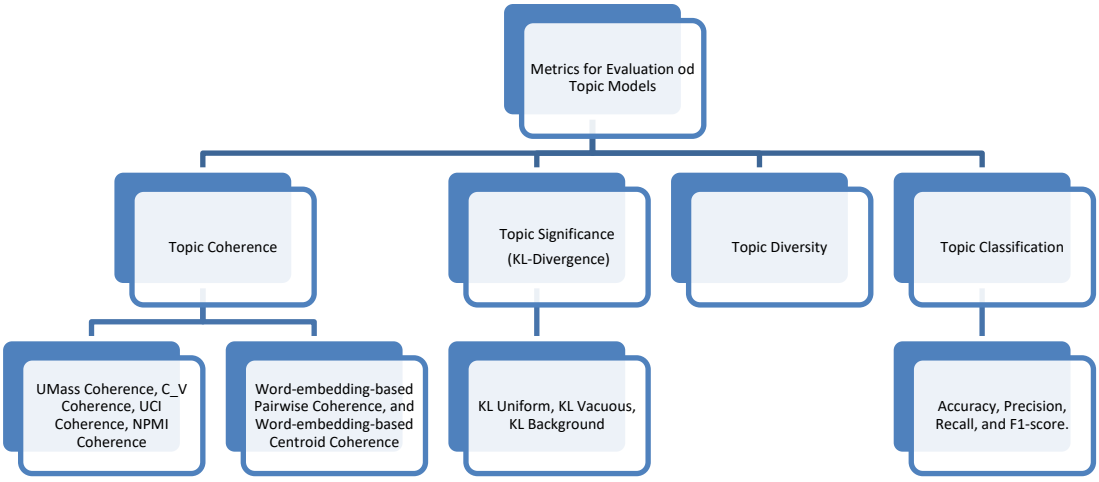
Metrics for Evaluation od Topic Models
Topic Coherence
Topic Significance (KL-Divergence)
Topic Diversity
Topic Classification
UMass Coherence, C_V Coherence, UCI Coherence, NPMI Coherence
Word-embedding-based Pairwise Coherence, and Word-embedding-based Centroid Coherence
KL Uniform, KL Vacuous, KL Background
Accuracy, Precision, Recall, and F1-score.

Figure 2: Evaluation Metrics for Topic Models (Source: Self)

1.4. Hyperparameter Tuning

When employing any of the topic models, it is crucial to determine the optimal number of topics extracted by the model, as this significantly impacts the efficacy of the results. For topic models, the primary focus to achieve this objective is on using evaluation metrics to assess topic quality and structure. The metric for determining the appropriate number of topics was devised (Arun, et al., 2010). The research demonstrated richer topic structure extraction by interpreting LDA as a matrix factorization method, with a lower value of a proposed divergence metric between the factorized document-topic and topic-word distributions. Another framework integrates the concept of topic density for model parameter estimation to determine the best LDA model (Cao, et al., 2009). This work was based on finding the model's best 'K' using an iterative process over the correlations in the document collection independent of the size of the data set. The optimal value of hyperparameters can also be devised utilizing an array of evaluation metrics. Apart from determining the optimal value of K using proposed metrics that assess topic structure, alternative techniques focusing on the automatic inference of the optimal value using assumed priors have also been proposed across the literature.

1.5. Topic Modeling for Images

Owing to the success of probabilistic topic models for textual corpora, topic models have also been employed over image corpora for solving a wide variety of problems such as image annotation (Feng & Lapata, 2010) (Zhu, et al., 2018) (Zia, et al., 2019), image classification (Pan, et al., 2018) and content-based image retrieval (Aouadi, et al., 2015). Major approaches extend topic models to images by generalizing the notion of images as a bag of visual words from which latent topics may be inferred. Further, the inference procedure can be combined in a multimodal setting with associated textual data assumed to be drawn from a standard distribution to enhance the classification results further. Using topic models directly over image data and accompanying textual data makes them quite useful for image annotation and content-based image retrieval since both problem domains operate with multimodal text and image data.

2. Literature Review

Researchers have made significant strides in topic modeling, image captioning, and image classification. Researchers have developed metrics and techniques for assessing topic model quality, optimal topic numbers, and image captioning quality. In image classification, evaluation methods measure accuracy. Topic modeling uncovers latent themes in unstructured data, finding various applications, including image captioning and ongoing exploration in image classification. Some of the work has been discussed below:

2.1. Topic Models

In recent times, neural techniques have been combined with topic modeling. In (Wu et al., 2023), ECRTM is introduced as a novel neural topic model to address "topic collapsing" issues using Embedding Clustering Regularization (ECR). ECRTM, with ECR, generates diverse and coherent topics and high-quality document-topic distributions. The authors introduce ECR, integrate it into ECRTM, and perform experiments to demonstrate its effectiveness. However, the study lacks internal metrics (Topic Classification, Topic Significance) for model quality on unseen data and external evaluation metrics specific to assessing topic models, limiting their alignment with external information.

The researchers in (Sadigov et al., 2023) employ sentiment analysis and topic modeling techniques, specifically using the LDA algorithm. A long-term memory-based sentiment analysis model with word2vec embedding is utilized to gauge Twitter users' sentiments toward distance education. An LDA-based topic model is also constructed to identify and analyze the topics discussed within the Twitter dataset. However, the study doesn't delve into the quality of topics generated by LDA. Topic quality, coherence, and interpretability are essential aspects of topic modeling.

As the applications of topic models for exploring extensive document collections expand significantly, challenges of comparing the available algorithms proposed across literature, given the diversity of datasets and criteria utilized by researchers, and the selection of a suitable metric to evaluate results arise comprehensively. Rüdiger and others (Rüdiger et al., 2022) have conducted a novel in-depth study on the performance of topic models and quality

evaluation metrics. Based on their research, the authors aim to determine the criteria for selecting the best topic model for any application and how effective hyperparameter tuning, especially for the value of the number of topics, K, may be achieved using evaluation metrics proposed across the literature. In their study, the authors conduct a two-phase evaluation: a comprehensive examination of the topic modeling algorithms to determine the most suitable performant model and a thorough evaluation of the quality metrics to assess the interplay of performance, hyperparameter selection, and metric results. The aim is to identify candidate metrics and algorithms yielding the best clustering accuracy. The former is achieved through a holistic comparison of commonly deployed, generic topic modeling algorithms and an in-depth assessment of their relative performance. The distinction is made against ground-truth clustering, enabling an unbiased evaluation of results. For the latter, the authors analyze the correlation between existing metrics and ground-truth clustering to determine the suitable conditions for effectively utilizing these algorithms. In the study to determine the best topic model, the authors focus on the accuracy of cluster assignment by comparing prominent topic assignment for documents to a known clustering using metrics such as Adjusted Rand Index (ARI) and Adjusted Mutual Information (AMI). These metrics are compared to previous studies' intrinsic evaluation of interpretability and coherence as the authors believe that the internal cluster validity measures are insufficient to justify the performance of algorithms due to the unavoidable problem of synthesizing seemingly coherent but nonsensical topics. The resultant external cluster validity measures, selected by the authors, are necessary for a holistic evaluation. For this part of the study, the authors have evaluated the performance of various non-probabilistic and probabilistic models such as LSI, NMF, PLSA, and LDA, along with different inferential methods of LDA such as the online variational inference method (Hoffman et al., 2010) commonly employed in Gensim (Rehurek & Sojka, 2011) and Scikit-Learn (Pedregosa et al., 2011) and the collapsed Gibbs-Sampling method (Porteous et al., 2008) employed in the implementation in Mallet (McCallum, 2002). Evaluation over a curated dataset of English Wikipedia articles as documents and corresponding category information as ground-truth labels yield the Gibbs-Sampling LDA method to be dominant for large values of K over 20 and NMF for the small values of K for both clustering efficacy and scattering of results, consistent with results observed in existing studies. It was observed that an increase in the value of K negatively impacts performance. In the latter half of the study,

the performance of topic quality metrics concerning optimal clustering results is evaluated. The performance is evaluated solely based on cluster memberships using external cluster-validity metrics.

Commonly utilized metrics such as Perplexity, Topic Distance, Symmetric KL-Divergence, Topic Coherence, and Topic Divergence measures are included for evaluation. In the evaluation, the authors examine the correlation of quality metrics to the performance metric ARI and, based on the irregularity of correlations for all metrics across varying algorithms and hyperparameters, conclude that no suitable internal quality evaluation metric exists. Using such metrics and interpreting their results is subject to the domain-specific application; only for specific algorithms like Gibbs-Sampling LDA, metrics like perplexity exhibit stable correlation with the target metric. However, determining the optimal number of topics is subject to interpreting the results, and only an approximation can be achieved based on internal metrics. The proposed methodology does not evaluate the novel quality evaluation metrics such as Normalized Absolute Coherence, Topic Score, or Combined Coherence Values. Further, the authors do not evaluate novel topic modeling algorithms based on Deep Learning methods such as Top2Vec or BERT Topic, which may perform better than statistical methods. Lastly, the study did not include the evaluation of Bayesian optimization techniques based on prior estimations and their relation with the external cluster validity metrics, which can provide helpful insight into the efficacy of the metrics against manual hyperparameter tuning strategies.

The research work by (Qiang et al., 2022) discusses the application of topic modeling algorithms over short text data. This paper gives a comprehensive review of short-text topic modeling techniques. It discusses different methods based on Dirichlet multinomial mixture (DMM), global word co-occurrences, and self-aggregation. The state-of-the-art methods are evaluated on real-world data sets and have been compared with each other and long-text topic modeling algorithms. Short texts like news headlines, status updates, etc., infer latent topics that are discriminative and coherent, which is essential as it requires a semantic understanding of the corpus. Long-text topic modeling algorithms cannot solve this problem as they are based on word co-occurrences, which are in acute quantities across short text data.

The authors developed the first open-source library for short-text topic models, STTM, for use in Java that integrates the algorithms the author experiments with in one interface and benchmark datasets. The authors evaluated the classification, clustering, and topic coherence approaches over six publicly available data sets. Though this paper provides a comprehensive insight into the short text topic modeling algorithm, a few research gaps still exist. This includes a lack of effective standard topic model evaluation metrics. Topic coherence cannot discern the differences across topics, and established metrics can only assess a part of topic models. This paper visualizes topics by listing the most frequent words per topic. However, how to visualize an entire document using topic models is not discussed. In the experimental results, the performance of different methods is specific to the choice of the dataset. Hence, the proposed model lacks in predicting which method to use for the new corpus.

A general evaluation metric was developed to improve the topic model evaluation (Gan & Qi, 2021). The authors evaluate their proposal over standard datasets such as 20 Newsgroups, WOS 46985, and AG News and evaluate multiple variants of LDA based on online variational inference, collapsed Gibbs-sampling inference, and an LDA model based on GloVe word-embeddings and other topic models such as HDP. Their comprehensive metric focuses on judging the quality based on stability, predictive ability, and topic isolation of topic models. It combines the information given by perplexity, isolation (based on Kullback-Leibler or Jensen-Shannon divergence), stability (based on ranking by average Jaccard index), and coincidence (based on coherence measurements) into a topic score to determine the best value. Based on the evaluation of previous studies, the authors have found existing metrics to perform poorly and lead to non-optimal hyperparameter selection for domain-specific datasets. To mitigate this, the authors attempt to identify the characteristics of a reliable topic model: good predictive capability, which is identifiable by low perplexity; high isolation between different topics to avoid redundancy, which is identifiable by topic diversity and similarity metrics; high stability to ensure repeatability which is indicated by high topic stability metric values, and unique directionality and reliability of topics, a consequence of high coincidence which is characterized by coherence measures. As each metric evaluates specific characteristics of a topic model and all factors are reliable to some

extent, a standard metric that combines the individual metric scores into a typical score is proposed as shown in equation 1.

$$Topic\ Score = \frac{Perplexity \times Coincidence}{\sigma_{JS} \times Stability} \quad (1)$$

The authors propose a methodology for identifying the best number of topics for LDA models by calculating the extent of model stability across different values of K. The method involves generating an initial topic model over the complete dataset, randomly selecting samples from the dataset to create multiple topic models, and computing the coincidence and stability of each model against the test model. The remaining samples are used as the training set, and the isolation and perplexity for each test sample are computed. Topic superiority scores of sample models are then averaged across varying values of k, and the minimum value of k whose sample models have a significant average topic superiority score is selected as the optimal value. Conducted experiments over different chosen datasets indicate the effectiveness of the novel metric. However, this is lacking in evaluating the proposed metric over domain-specific datasets to determine the extent of generalization. Other aspects include the limited scope of evaluation metrics where only standard metrics have been considered and recently proposed metrics such as Normalized Absolute Coherence have not been compared.

The authors (Albalawi et al., 2020) discuss discovering helpful information or the topic discussed in short text data. They have examined and compared five topic models for short-text data, which include Latent Semantic Analysis (LSA), Latent Dirichlet Allocation (LDA), Non-negative Matrix Factorization (NMF), random projection (RP), and principal component analysis (PCA) over two selected datasets. These models were evaluated using metrics assessing topic quality and over metrics such as recall, precision, F-score, and topic coherence. Over these standard statistical evaluation metrics, the results look promising for LDA and NMF. As per the paper, these two topic modeling methods delivered more meaningful extracted topics. The biggest challenge with inferring topics from short text data arises because of the significantly low number of topics per document and noise, leading to incorrect topic discovery. Overall, the comparative analysis of topic modeling methods for

short text data is an active area of research, and no single best method suits all situations. Therefore, it is essential to carefully select the most appropriate method for the specific dataset and research question.

The research study by authors (Kherwa & Bansal, 2019) comprehensively reviews topic modeling. The author discussed several versions of topic models, including classification hierarchy. Latent Semantic Analysis (LSA) (Deerwester et al., 1990), Latent Dirichlet Allocation (LDA) (Blei et al., 2003), Probabilistic Latent Semantic Analysis (PLSA) (Hofmann, 1999), and Non-negative Matrix Factorization (NMF) (Lee & Seung, 1999) are some of the most commonly employed topic models. Depending on the algorithm's nature and the inference method, these can be broadly classified into probabilistic and non-probabilistic. The paper discusses posterior inference techniques, LDA's evolution models, and their applications in different technology areas. The report concluded with detailed discussions on the challenges of topic modeling. The paper provides a good insight into different topic modeling algorithms and challenges. The empirical evaluation was done for only two algorithms, LSA and LDA, over two datasets, while other algorithms were not considered for evaluation. The hyperparameter 'K' was chosen as 10 randomly for the experiment setup. Proper evaluation metrics should have been used to determine the optimal hyperparameter 'K.' The paper discusses the evolution of LDA based topic models but does not provide insight into evaluation metrics to evaluate hyperparameters for these models.

A study by Chen and others (Chen et al., 2019) focuses on evaluating the popularly employed topic modeling algorithms over short-text datasets to compare the effectiveness of both algorithms. Also, they determine the most suitable algorithm for such datasets, which can be beneficial for understanding latent topics from short texts such as those found in streaming social messages. The authors comprehensively conducted experiments over several publicly available short-text datasets and found that NNMF performs better than LDA in primary settings. They propose a model called "Knowledge-Guided Non-negative Matrix Factorization for Better Short Text Topic Mining" (KGNMF). It utilizes external knowledge for semantic regulation, with low-rank formalizations, implying a time-efficient algorithm. The experiments conducted over representative corpora demonstrate the effectiveness of

KGNMF in short-text topic mining. The experiments conducted by the authors are divided into two parts. The authors compared the core LDA and NMF models with different experimental hyperparameters over several public short-text datasets in the first part.
The datasets included news headlines, tweets, and product reviews. They evaluated the models based on their ability to generate meaningful topics and their coherence using metrics such as perplexity, coherence score, and topic diversity. The results showed that NMF performs better than LDA in basic settings on the news headlines and tweets datasets. LDA performed better than NMF on the product reviews dataset. The authors attributed this to the sparsity of the product reviews dataset, which may have resulted in better performance for NMF. In the second part, the KGNMF model is proposed, and experiments over representative corpora against typical short text topic models are conducted to demonstrate the effectiveness of KGNMF. However, there is a lack of comparison using novel quality evaluation metrics. The model was not evaluated against deep-learning topic modeling techniques or external quality assessment metrics to compare the topic assignments with ground-truth clustering. This may be vital to determine the practical efficiency of the chosen algorithm against others.

Latent Dirichlet Allocation (LDA) is the most prominently employed topic modeling technique. However, it is subjected to order effects when running on different datasets, i.e., if the training order is shuffled, the algorithm generates different topics. This effect can lead to fallacious results, inaccurate topic descriptions, and reduced efficacy of text classification results. This paper (Agrawal et al., 2018) proposed a method to generate more stable LDA distributions and how it can be used for further analysis. Different implementations of LDA were carried out for different training strategies (VEM, Gibbs Sampling) and evaluated via classification accuracy and topic stability. The author used Differential Evolution (DE) to fine-tune the hyperparameters of LDA, using a tool proposed with the name LDADE. The model is evaluated on different datasets, including Stackoverflow, SE papers, and defect reports for software from NASA. The result shows that the core LDA demonstrates considerable topic instability.
With tunings from LDADE, cluster instability can be reduced, leading to improved results across supervised and unsupervised learning. This paper is an empirical study inclusive of

many biases that can affect the final results. These biases include sampling bias, learner bias, evaluation bias, and order bias, hence the research gaps for the paper. The experiment uses preprocessed data sets; thus, results can vary (wrong or less improved results) if data sets are pre-processed differently (sampling bias). The optimal set of hyperparameters is not taken into account while running LDA. This can provide a significant improvement in the results (learner bias). This paper uses only topic similarity and F2 metrics for evaluation, but other robust metrics, like perplexity, coherence, etc., are used to assess the performance of stable LDA (evaluation bias). The random sampling method picks up data from the data set. This paper is to study the order effects; hence, the random sampling method might not be the most suitable. The method is not validated on different data sets from different domains, the use of which may lead to the invalidation of the claims.

2.2. Hyperparameter Tuning: Optimal Number of Topics 'K'

Recently, authors (Weisser et al., 2023) explored the challenges posed by short and sparse social media texts, such as Twitter, for the conventional Latent Dirichlet Allocation (LDA) topic model. It compares LDA with two models designed for sparse data: the Gibbs Sampler Dirichlet Multinomial Model (GSDMM) and the Gamma Poisson Mixture Model (GPM). A novel evaluation method involving pseudo-document simulation is introduced. In a case study involving Covid-19-related tweets, the study finds that standard coherence scores, commonly used for topic model evaluation, are ineffective as evaluation metrics. The proposed solution still necessitates separate hyperparameter optimization for the individual models, employing coherence scores or recommended priors from existing literature.

This research work (Lewis & Grossetti, 2022) addresses concerns about determining the optimal LDA configuration by introducing a set of parametric tests based on the multinomial distribution of the original LDA. The proposed methodology defines a set of statistical procedures that identify and evaluate the optimal topic model. It is based on the conjecture that if 'K' is the minimal number of topics that define that corpus, then these 'K' topics should define the corpus better than any other specification. More than 'K' should not give any substantial value addition. The methodology was evaluated on the U.S. Presidential Inaugural Address Corpus. The proposed method uses a chi-square test followed by topic

stability determination to determine its effectiveness in those settings. Again, two more tests are conducted to address the overall model adequacy using a series of Chi- square tests.
The proposed approach was compared with the perplexity index, which outperformed significantly. The result showed that the 92 topics describe the corpus mentioned above. The results are validated through a simulation study and compared to standard heuristic metrics like the perplexity index. The proposed method was also validated for topic stability. Topics were segregated into informative and uninformative categories. Also, it has been observed that uninformative topics do not have significant explanatory power. The proposed methodology is evaluated only on one data set, not the different data set that varies in size and domain. Hence, there is a lack of validation for different data sets. Secondly, the optimal 'K' is compared using the perplexity index only, with many other metrics to do the same. The optimal 'K' from the proposed methodology must be compared with other heuristics metrics, such as coherence or diversity metrics. The proposed model relies only on one metric for comparison. Also, the author did not discuss the accuracy of results on short text data.

The results of various experiments on the model fitting problem were presented (Sbalchiero & Eder, 2020). They assessed its accuracy over long-text datasets. This article discusses the relationship between text length and the optimal number of topics. Their research gives a mathematical model for the same. The study was conducted via three experiments applied to text segments of varying sizes. The text length varied to explain the change in the best number of topics to detect in the corpus. The experiment was performed on a large corpus of English novels, then on the reduced corpus of 10 novels, and lastly on an equal number of text chunks. The author calculated the log-likelihood for each sample size to identify the best number of topics. The results were stable up to some text length and then inferred optimal topics decreased when the size of the text chunk increased. The results show a clear relationship between the optimal number of topics 'K' and the size of the sample text, hence presenting a Sbalchiero-Eder rule. Accordingly, the optimal number of topics is inversely proportional to the text segment length, i.e., larger text portions corresponding to a lower value of K. The equation is presented as this: $y = f(x)$, where y denotes the best number of topics and x denotes the size of the text chunks in word tokens. The function f is a steep

decreasing function that may be a power or logarithmic model. The theory is based on log-likelihood and validated over only one data set. The data set that is being used is of literary novels, and results may vary for some other datasets, like scientific paper datasets, as they are denser in content. However, the author did not discuss any comparison with other evaluation metrics for optimal 'K' like perplexity, coherence, and diversity.

Recent work by Peikert and others (Peikert, et al., 2021) analyzed and proposed some objective functions for evaluating topic models to determine optimal hyperparameters. Empirical studies conducted with the NMF algorithm using the proposed objective functions over different datasets yield promising results for determining the optimal values of the number of topics. The objective functions used in the approach improvise over the topic coherence metrics proposed by Röder et al., 2015, which are used for internal topic evaluation. These are combined with external topic evaluation functions, which evaluate topic similarity across pairs of generated topics. The combination allows for a holistic assessment with an extended focus on interpretability. The external topic evaluation measure focuses on the interrelationship between topics and integrates this into a scoring. Based on the observation that increasing the value of K leads to the factorization of coherent topics into constituents that still resemble the parent topic, a mapping $\vec{m}$ Of every topic to its most similar topic, the result is generated using either equation 2 or 3:

$$\vec{m} = \overrightarrow{m_{ove}} = \begin{pmatrix} \max\left\{\frac{|t_1 \cap t|}{2N} \,\middle|\, \forall t \in T \setminus \{t_1\}\right\} \\ \vdots \\ \max\left\{\frac{|t_K \cap t|}{2N} \,\middle|\, \forall t \in T \setminus \{t_K\}\right\} \end{pmatrix} \quad (2) \qquad \vec{m} = \overrightarrow{m_{cos}} = \begin{pmatrix} \max\left\{\cos_{t_1,t}(\theta) \,\middle|\, \forall t \in T \setminus \{t_1\}\right\} \\ \vdots \\ \max\left\{\cos_{t_K,t}(\theta) \,\middle|\, \forall t \in T \setminus \{t_K\}\right\} \end{pmatrix} \quad (3)$$

The mapping is combined with the topic coherence of individual topics to obtain a measure $\overrightarrow{\phi_{C_{C_v}}}$. This measure assesses topics over both coherence and the presence of unique features. The measure is then averaged to obtain the combined coherence value C_{C_v} as shown in equation 4. To compute C_{C_v}, $\overrightarrow{\phi_{C_{C_v}}}$ is calculated by equation 5.

$$C_{C_v} = avg(\overrightarrow{\phi_{C_{C_v}}}) \quad (4)$$

$$\overrightarrow{\phi_{C_{C_v}}} = \overrightarrow{C_v} \odot (\vec{1} - \vec{m}) \quad (5)$$

Using the above functions, a high-level procedure for tuning the number of topics, K, is devised as follows: firstly, topic models for varying values of K are created. For each model, the topic coherence value C_v is computed, along with the vector mapping for topic similarity among pairs of topics, $\vec{m}$. These are combined and aggregated to obtain a combined coherence value C_{C_v}. The topic model with the maximum C_{C_v} value corresponds to the best value of K. This work lacks in exploration of possible improvements using metrics for comparing topic similarities, such as Jaccard similarity and word-embedding-based similarity measures instead of overlap or cosine distance. The proposed functions differ from existing approaches, such as the metrics proposed by Juan et al., 2009, Arun et al., 2010, etc., and whether the procedures apply in domain-specific applications.

A novel method for selecting LDA hyperparameters based on a hierarchical Bayesian framework involving setting priors on the hyperparameters was developed recently. The method uses variational inference for estimating the posterior distribution of the involved hyperparameters. Based on the Monte-Carlo Markov Chain, (Xia & Doss 2020) is a fully Bayesian method for generating the empirical Bayes estimate of the number of topics. The proposed method is scalable, and experiments conducted by the authors demonstrate its ability to handle large datasets using stochastic optimization techniques. The proposed method evaluated several existing methods for hyperparameter selection in LDA. The evaluation shows that the method outperforms current topic coherence and diversity approaches. The proposed method is deployed over an empirical dataset of scientific publications to extract topics and analyze research trends. The results show the identification of meaningful topics and insights into research trends in the field. The method was evaluated using only topic coherence and topic diversity but not with other recently proposed metrics. There is a lack of comparison of the approach's effectiveness with other optimization techniques and the feasibility of the process in practical applications in terms of computational efficiency. Future studies could explore other evaluation metrics, such as semantic similarity, interpretability, or predictive performance, for a comprehensive and extensive method evaluation.

A method was proposed by (Hasan et al., 2020) for devising the most appropriate (optimal) number of topics for LDA, which discusses the improvements in perplexity and coherence metrics. They present a normalized approach that uses the coherence score to determine the degree of semantic similarity between words in a topic. The normalized coherence score is calculated by dividing the coherence score of a particular number of topics by the maximum coherence score across all tested numbers of topics. The authors argue that this normalization step helps to reduce the influence of the size of the corpus on the coherence score and makes it easier to compare models trained on different datasets. They compare their approach to several other methods for selecting the number of topics, including the perplexity measure, the topic coherence measure, and the topic uniqueness measure for different datasets. The research focused on evaluating the extent of the influence of perplexity and coherence over the quality of topics resulting from the LDA model to improve the F-Measure for a topic classification task. They found that the best topics were determined based on values of the normalized absolute perplexity (NAP) and normalized absolute coherence (NAC) (UMass) metrics. The results show that the normalized coherence score method outperforms the other ways in terms of both coherence and interpretability of the topics. The proposed method was neither tested on larger data sets to assess its generalizability nor compared with more recent approaches and other state-of-the-art methods, such as coherence-based or non-parametric Bayesian methods. The paper normalizes the coherence score by dividing it by the maximum coherence score across all tested numbers of topics. However, it is unclear how the performance of the proposed approach would be affected by different normalization strategies or the use of unnormalized coherence scores. The comparison with other normalization methods, such as scaling or standardization, is also missing, which may impact topic modeling accuracy. While the paper focuses on LDA, it would be valuable to explore whether the proposed normalization method could also be applied to other topic modeling algorithms (such as Hierarchical Dirichlet Process (HDP) or Non-negative Matrix Factorization (NMF)) or not.

Non-evaluation techniques relying on Bayesian optimization have been recently proposed to automatically determine the best hyperparameter settings (Chen & Doss, 2019). A Bayesian mixture model that controls the sparsity of the topic assignments using additional parameters

to estimate the number of topics was presented. The methodology imposes a prior over K to determine the optimal value during inference. Consequently, the distribution of latent topics is a mixture of distributions on spaces of varying dimensions, for which estimation is achieved via a proposed modification of the Metropolis-Hastings algorithm. The authors evaluate the proposal in terms of performance on several datasets, including the 20NewsGroups dataset and the NIPS dataset, and with other methods, such as the harmonic mean method, the Laplace approximation, and the variational Bayes method, for estimating the number of topics for the LDA model. The results showed that the proposed model outperformed the other methods regarding accuracy and robustness. The proposed model could accurately estimate the number of topics, even in cases where the data was highly imbalanced or the valid number of topics was unknown. However, the model was not evaluated on real-world extensive data sets with millions of documents and thousands of topics from different domains and languages. The paper focuses on the problem of estimating the number of topics in the LDA model; however, the issue of topic interpretability is not addressed. There is a lack of comparison with deep learning-based approaches, such as neural topic models, and how the proposed methodology can be extended to other topic models, such as the Hierarchical Dirichlet Process (HDP) and the Correlated Topic Model (CTM).

2.3. Image Captioning using Topic Modeling

An unsupervised approach for accelerated image annotation is also proposed by (Wang et al., 2019). Their proposed methodology aims to organize unsupervised images using latent features extracted from topic models. Their experiments demonstrate the technique's usefulness for image annotation and retrieval tasks without sacrificing labeling accuracy. In the approach to extract latent features, the authors propose a three-stage pipeline: First, a feature extraction stage that utilizes unsupervised CV techniques such as Scale-Invariant Feature Transform (SIFT) or an internal representation from activation channels of a pre-trained Convolutional Neural Network (CNN) to synthesize relevant features from an image in a high-dimensional space. Secondly, a clustering stage utilizes an unsupervised clustering algorithm such as K-Means to discretize extracted features for a lower dimensional representation, creating an equivalent visual ‘bag-of-words’ model for every image

document. Lastly, latent topic representations are extracted from image documents using discretized features via a topic modeling algorithm such as Latent Dirichlet Allocation (LDA). Using unsupervised clustering and topic modeling algorithms allows streamlining the annotation and querying process even in the absence of ground truth image labels and hinted structure and demonstrates better performance compared to baseline approaches such as PCA K-Means and Variational Auto-Encoders. The results show that using latent topic embeddings, the proposed unsupervised image clustering pipeline can group topics of similar ground truth labels. The model is highly dependent on feature embeddings, which can hamper the performance if not appropriately considered. The qualitative and quantitative evaluation results showed that the feature extraction hyperparameters could affect the results in the pipeline's clustering and topic modeling stages. However, there is a high dependency on feature embeddings and feature extraction hyperparameters, which can impact the clustering and topic modeling results. The model is evaluated using the KL-Divergence metric for quantitative evaluation to find hyperparameters and global averaged KL divergence for clustering. No other metrics have been considered. The proposed model is evaluated on a balanced data set and not tested on skewed data sets. No study has been conducted to assess the model's efficacy on human annotation on preprocessed datasets.

Another method for image captioning was proposed by (Dash et al., 2019), which imparts image topic objects' presence and relative orientation as vital information for driving the caption generation process. This work extended over recent work on image captioning using topic models, such as the work done by (Zhu et al., 2018). They aimed to emphasize the image context in the model to imitate the human approach and filter out irrelevant attributes from the captions. The proposed model followed an encoder-decoder framework; unlike sequential models, it followed the merge model framework, merging the multimodal features before decoding. The model consisted of two components. The first identifies the image context based on the topic by utilizing image features and mapping them to topic models. In the second component, the topic information is used to generate captions. Topic extraction was performed using LDA from the caption data associated with training images. These topics were used to train a neural network classifier to predict topics for a given image, with topic words composed of object categories. The decoder network includes an LSTM

network, using the topic information, image features, and the current incomplete caption as input. The network then merges the output of the encoded captions and the image features to predict the next word for the caption.

The method is evaluated over benchmark datasets: Flickr8k and COCO. The results demonstrate the approach can outperform several state-of-the-art image captioning models on various evaluation metrics, including BLEU, METEOR, and CIDEr. In addition, the authors conduct a human evaluation, showing that their proposed method generates more informative and diverse captions than other methods and has potential applications in various domains, including image and video captioning and visual question-answering. The paper focuses on generating individual captions for images, but there is a need to develop a coherent and consistent storyline or paragraph for multiple images in a sequence. The proposed method uses the older models (InceptionV3, LSTM) instead of more promising transformer-based models and does not compare it with transformer-based annotation methods. The model is not evaluated on noisy and unannotated real-world image data sets. Future studies could investigate how to generate long-term coherent captions for multiple images using the proposed method.

The authors (Zia et al., 2019) proposed a framework for generating topic-sensitive descriptions for images that incorporates topic-sensitive word embeddings into the captioning process, developed by combining associated topics with word embeddings. Their methodology uses HDP for topic identification and R-CNN for image feature extraction. The proposed model grasps the semantic relation and polysemantic words that define the images and synthesizes apt descriptions for the target images. The proposed model is evaluated using BLEU-N, METEOR, and CIDER scores on the state-of-the-art captioning datasets Flickr8K, Flickr30K, and MSCOCO. These evaluation metrics measure how closely the generated caption matches the human-annotated images' five captions. The proposal significantly improves across most evaluation metrics except BLEU-4 against existing state-of-the-art models on selected datasets. The proposed model is compared further to measure the performance of the generated topic-contextualized description of images against state-of-the-art models. The result demonstrates that the model can generate sensible and more relevant image descriptions. The proposed model lacks in evaluation using better image feature

extraction methods (YOLO, Faster R-CNN, etc.) The model uses RNN instead of transformers for caption generation from topic-sensitive word and image embeddings.

The authors (Zhu et al., 2018) proposed an attention mechanism for automatic image captioning by integrating image topics. Existing models rely on attention to identify relevant image features based on feedback from a caption generator. However, the approach is constrained by the absence of direct, high-level guidance from the image for relevant features. The proposed mechanism integrated image topics into the attention model to guide the selection of relevant image features. The model network followed the encoder-decoder architecture, in which the encoder obtains image features, and the decoder interprets the features to map them to captions. The encoder architecture was comprised of a feature extractor composed of a deep-CNN model for extracting visual features, a topic predictor for which a given image caption utilizes LDA (over the caption data to generate image-topic pairs used for training a single label classifier) to classify unseen images into topics, and an attribute vector predictor for identification of most salient terms for an image (through a multi-label classifier trained over an attribute vocabulary built from supervised caption data). The results from the feature extractor, the topic predictor, and the attribute predictor were passed to topic-guided spatial and semantic attention models to determine the most relevant visual attributes based on the inferred topic and feedback from the decoder network. The decoder employs an LSTM network to generate image captions based on the feedback on the features and attributes provided by the encoder.

The method outperforms several state-of-the-art image captioning models on benchmark datasets. It has potential applications in various domains, including image and video annotation/captioning and visual question-answering. The paper evaluates the proposed method on the MS-COCO dataset with limited to no exploration of the effects of hyperparameter tuning on model performance. However, evaluating the model on more extensive and diverse datasets is essential to assess its scalability. The article lacks the interpretability of results as it does not provide a detailed analysis of how the topic model can extract relevant topics from the images. There is inadequate modeling of semantic relationships in the proposed model, and it's been suggested as future work in the paper to merge spatial and semantic attention components in a single network.

The proposed method was evaluated on the image captioning task using various automatic evaluation metrics without human evaluation of the generated captions. Further, the proposed method was evaluated under controlled conditions, where images are well-defined and annotated. However, in real-world scenarios, images can be noisy and unannotated. Hence, it is essential to investigate the robustness of the proposed method to such conditions. Also, it is unclear whether the proposed method can be generalized to other applications, such as visual question-answering or image retrieval.

2.4.Image Classification using Topic Modeling

The authors (Tsapatsoulis et al., 2022) explored using pre-trained deep learning models, such as Resnet-50, for transfer learning, including image retrieval. They previously introduced topic modeling for Instagram photo retrieval using hashtags. This study compares content-based image classification through transfer learning with hashtag-based topic modeling across 24 concepts. The study lacked specific evaluation metrics for topic modeling, hindering its quality and effectiveness assessment.

The paper by (Miller et al., 2020) improves image classification by incorporating natural language understanding with image metadata, using descriptions, titles, and tags through a multi-modal model. It innovatively applies natural language understanding and transfer learning to external image-related features. The study does not incorporate topic modeling.
A topic network model was proposed by (Pan et al., 2018), combining topic models with deep CNNs for image classification. Their approach modifies the standard architecture of CNNs by replacing pooling layers with topic models to prevent information loss due to pooling, allowing images of arbitrary sizes to be fed to the network, relaxing the fixed size constraint, and preventing information loss due to cropping and warping to resize the images. Their approach replaced the pooling layers in a CNN architecture with the LDA and Markov topic random fields (MTRF) topic model to create a topic network, as the convolutional layers were independent of the image sizes. The topic model treated variable-sized images as documents, with features identified by the previous convolution layers forming the visual words of which the image is composed. The vocabulary of the corpus was built by applying clustering over all the features of the images. Following the mapping of images to visual

words, LDA was used to learn the topic distribution. The result of the topic layer was a K-dimension vector, with K equal to the number of topics, representing the distribution of topics in the image, which was then fed to the fully connected layers for classification. The proposed architecture outperformed standard image classification datasets such as Caltech101 and VOC2007. Both the topic models, LDA and MTRF, performed well in image classification with the original size. Compared to other approaches, it's been shown that the method is agnostic in the choice of topic model for the topic layer, making the proposed framework flexible and extensible. Although the model looks very promising, there is a lack of comparison against more sophisticated deep learning methods such as image transformers. The model is not evaluated to understand the effects of hyperparameter tuning on the results.

Another topic network model was proposed by (Pan, et al., 2018), combining topic models with deep CNNs for image classification. Their approach modifies the standard architecture of CNNs by replacing pooling layers with topic models to prevent information loss due to pooling, allowing images of arbitrary sizes to be fed to the network, relaxing the fixed size constraint, and preventing information loss due to cropping and warping to resize the images. Their approach replaced the pooling layers in a CNN architecture with the LDA and Markov topic random fields (MTRF) topic model to create a topic network, as the convolutional layers were independent of the image sizes. The topic model treated variable-sized images as documents, with features identified by the previous convolution layers forming the visual words of which the image is composed. The vocabulary of the corpus was built by applying clustering over all the features of the images. Following the mapping of images to visual words, LDA was used to learn the topic distribution. The result of the topic layer was a K-dimension vector, with K equal to the number of topics, representing the distribution of topics in the image, which was then fed to the fully connected layers for classification. The proposed architecture outperformed standard image classification datasets such as Caltech101 and VOC2007. Both the topic models, LDA and MTRF, performed well in image classification with the original size. Compared to other approaches, it's been shown that the method is agnostic in the choice of topic model for the topic layer, making the proposed framework flexible and extensible. Although the model looks very promising, there is a lack

of comparison against more sophisticated deep learning methods such as image transformers. The model is not evaluated to understand the effects of hyperparameter tuning on the results. An overview of the research study conducted is provided in Appendix Table 1.

2.5. Research Gaps

Although there has been rapid progress in deploying topic models over image data, many research gaps still exist, which may guide future research. Notable ones include:

- **Limited exploration of hyperparameter tuning:** Hyperparameters are crucial for the effectiveness of topic modeling over images. Optimal results can be achieved by tuning these hyperparameters effectively. There is still a lack of research on how to tune them effectively for optimal results. Researchers could investigate the impact of hyperparameters, including the number of topics, alpha, beta, and learning rate, on the quality of application-specific results.
- **Inadequate modeling of semantic relationships:** Most existing methods of topic modeling over images are limited in their ability to model semantic relationships between different visual concepts. Further research could focus on developing more sophisticated models that capture semantic relationships between concepts to improve application-specific performance.
- **Lack of consideration for spatial information:** Most current topic modeling approaches for images ignore the spatial layout of objects in the images due to the inherent assumption of the bag-of-words model. There is a need for research that incorporates spatial information in topic modeling to improve performance.
- **Limited exploration of topic coherence:** Topic coherence measures how semantically meaningful the topics are. Although topic coherence has been explored in text-based topic modeling, it has not been extensively studied in image-based topic modeling. Further research could focus on developing new methods for measuring topic coherence in image-based topic modeling.
- **Lack of research on the interpretability of topics:** Although topic modeling can produce meaningful clusters of visual concepts, it is still unclear how to interpret these clusters effectively, especially in images. Researchers could investigate new methods for visualizing and analyzing the topics to improve the interpretability of the results.

3. Problem Definition

This work aims to explore unstructured image datasets for caption generation, investigate topic modeling algorithms' efficacy, and develop a framework for image classification through topic identification using transfer learning and optimization algorithms to enhance accuracy and efficiency in real-time settings.

4. Objectives of Research

- To analyze unstructured image datasets to create captions and validate the existing dataset captions regarding accuracy, relevance, and quality.4
- To examine and explore the efficacy of various topic modeling algorithms for image captions (short text data) to identify the best topic model. Chapter 2
- To choose appropriate metrics for appraising topic models and devise the optimal number of topics and other hyperparameters. 3
- To develop a framework for image classification by utilizing topic information from the image caption dataset through transfer learning. 5
- To validate the accuracy of the proposed framework for image classification and optimize the same by utilizing optimization algorithms to decrease errors and boost effectiveness.

5. Research Methodology

This research aims to design an efficient model for image classification using topic models for the image data set. For the same, a methodology comprising five phases is proposed: Data Collection, Data Preprocessing, Image Caption Generation, Topic Model Evaluation, Hyperparameter Tuning, Image Classification via Transfer Learning, and Optimization. These phases allow the development of the effective framework required for precisely classifying images using semantic information embowed by latent image topics. Post optimization, the methodology further investigates the potential real-time applications that directly benefit or are enhanced via the proposed framework.

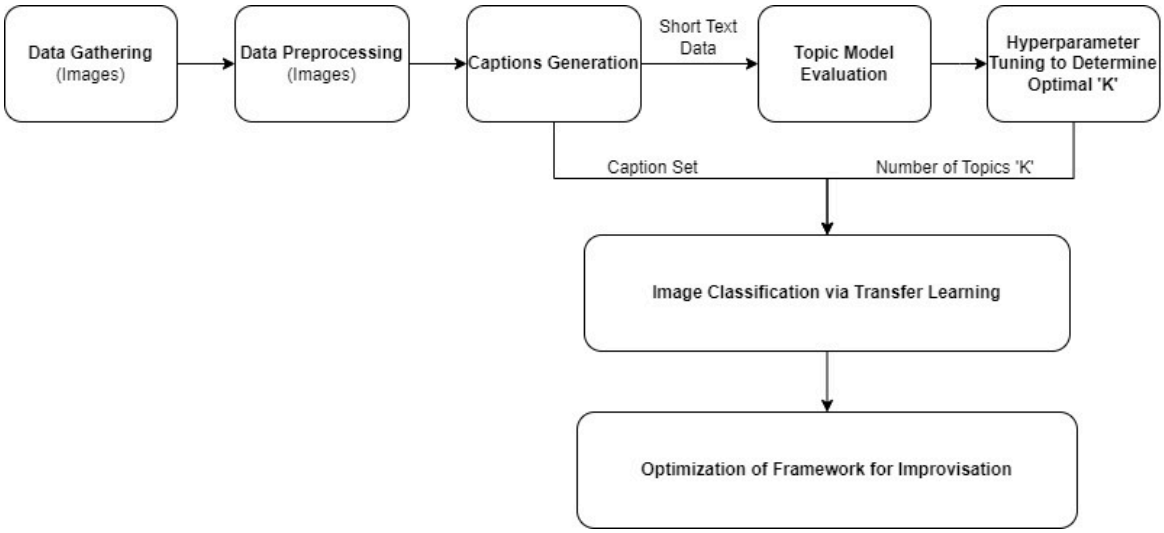

Figure 3: Overview of the Research Methodology (Source: Self)

Phase I: Data Gathering

The primary phase will involve selecting suitable datasets to develop a topic model that can infer underlying themes and extract the representations from images. For the same, an appropriate image dataset is required. Along with the image dataset, accompanying captions must also be present to develop the necessary framework to generate captions for images. One key consideration when selecting an image dataset for training a topic model is the size and diversity of the dataset. Other factors are the resolution and format of the images and any additional metadata or labels that may be available. For example, if the model is intended for image classification, including images with detailed labels or tags that describe their content may be beneficial. MS COCO, 2012, and Flickr8K, 2013 are suitably large datasets commonly employed in research for training image captioning models, which can be utilized to develop topic models.

Phase II: Data Preprocessing

Once the dataset has been curated, or an existing dataset is selected for training, the next phase comprises data preprocessing to remove irregularities and augment the dataset so that the trained model can generalize to unseen examples. Preprocessing would involve resizing images to a standard size for the image data. A pre-trained convolutional network architecture then processes the images to extract image features from hidden layers of the

model. Post-resizing, other processing techniques such as cropping, contrast stabilization, histogram equalization, gamma correction, rotation, shearing, and scaling may augment the dataset with more examples to force the network to learn general image features and prevent overfitting over the dataset. These image features constitute the corpora of 'visual words,' which a topic model can utilize to learn latent topics from images. For the corresponding caption data, text preprocessing strategies would be employed to sanitize the text of irregularities. These strategies would involve tokenization, stemming and lemmatization, stopword removal, and conversion to a bag-of-words representation. Common non-probabilistic and probabilistic topic models utilize the bag-of-words representation to represent documents in a corpus. The processed captions form the text corpora, which will be used to train a topic model and learn latent topics from captions.

Phase III: Image Caption Generation

The proposed framework encompasses a vital phase involving the generation of image captions for datasets lacking corresponding captions. The objective is to ensure the availability of captions that can be used with topic modeling algorithms to learn specific topics related to the image data. To accomplish this, deep learning models recently proposed in the literature are evaluated, and the most suitable model, based on performance in terms of metrics or human evaluation, is selected for the caption synthesis task. The synthesized captions are then pre-processed using the procedure described in Phase II for processing text data. Subsequently, these captions are utilized in Phase IV for topic modeling, similar to the pre-defined caption data found in image captioning datasets. By generating captions on the fly for datasets lacking such annotations, the framework offers generalization and effectiveness in handling real-world datasets where annotation is likely absent and human annotation is costly.

Phase IV: Evaluation of Topic Models and their Hyperparameters

Post-processing and synthesis of data, topic models over both image data are to be trained to extract image topics and the corresponding caption topics for a given image-caption pair. These topics will guide the classifier model employed to classify images into distinct labels. However, before a topic model can be employed, selecting the available non-probabilistic

and probabilistic topic models must be made to determine the most suitable topic model. Further, before topic models can be trained, several hyperparameters influence the training process, and the quality of results must be tuned. Only once a suitable topic model has been selected with the optimal choice of hyperparameters a classifier can utilize the extracted topics to classify the images. Topics produced by such a topic model are likely to be of high quality and would improve classification accuracy. Evaluation metrics such as topic similarity metric (proposed by Juan et al., 2009), diversity metrics (proposed by Arun et al., 2010), topic-coherence metrics (proposed by Röder et al., 2015), and other metrics for topic significance and performance over classification tasks are commonly employed to determine the effectiveness of topic models. These metrics provide an objective estimate of the performance of a model, which can be used to compare the model's effectiveness with other models. Of all topic models, the LDA model is commonly found across the literature to be the most performant, with the highest quality topics based on values of varying evaluation metrics. For improved efficiency, the hyperparameters of the topic models, notably the number of topics required by most non-probabilistic and probabilistic topic models, must be appropriately selected. The choice of these hyperparameters depends on the dataset over which the topic model is being employed. A poor choice of hyperparameters can adversely affect the performance of the topic model. The topic evaluation metrics are utilized to assess the performance of the topic model by determining the optimal hyperparameters. Based on the optimizing strategy of the metric, the best set of hyperparameters can be effectively determined. A collection of multiple evaluation metrics can be utilized to determine the hyperparameters. The agreement value of all metric objectives is suitable for the hyperparameter value.

Phase V: Image Classification using Transfer Learning

Once the topic model is selected along with the optimal choice of hyperparameters, the next phase involves training topic models using the processed dataset to learn latent image and caption topics. The multimodal classifier then utilizes these topics to classify images as per a set of distinct labels. A transfer learning approach utilizing the VGG16 model is used to produce image features, where feature representations from the last few hidden layers of the network are extracted to train topic models. The pre-trained model can be fine-tuned on a

new dataset of images to use transfer learning for learning image features that can be used with topic models relevant to the classification task. Once the topics are extracted, these are utilized by a classifier to determine the labels for a given image caption pair accurately. Using topics covering underlying latent information across both modalities will likely benefit classification, as the classifier can benefit from a high-level, low-dimensional data representation.

Phase VI: Performance Evaluation and Optimization

Once the framework for the classification task is developed and successfully trained, the final phase will evaluate its performance on classification and captioning tasks to assess its performance and determine its utility for real-time applications. For classification tasks, performance is evaluated using standard metrics of accuracy, precision, recall, and the harmonic mean of precision and recall, the F1-score. These metrics provide a quantitative measure of the accuracy and effectiveness of the classifier, allowing for the comparison of different models and the identification of areas for improvement. In addition to these traditional evaluation metrics, it may also be helpful to consider more specialized task-specific metrics such as mean average precision and mean average recall. Further, the framework is also evaluated over captioning tasks to determine the extent of generalization. This evaluation uses the topic model results from training a guided caption generator model to produce semantic captions based on information from the image and caption topics. These captions can then be compared with the dataset captions to determine whether topic model guided model captions are better at describing the image. Once assured of high quality, the classification results from the framework can be utilized to evaluate the effectiveness of associated dataset captions in describing the image. The same can be performed using only the caption topics to classify the images. Based on the results, the appropriate effectiveness and relevance of the caption in describing the associated image can be determined. Once the framework is successfully trained, strategies will be employed to optimize the process to enable effectiveness in a real-time environment.

6. Expected outcome of the research

The proposed framework aims to use a topic model on image captions to extract meaningful topics. Hyperparameter tuning will be employed to optimize the topic modeling algorithm, improving its performance and accuracy. The image classification system becomes more efficient and precise by utilizing topic information from the topic modeling and textual description from image captioning algorithms. The framework's effectiveness and resilience for image classification tasks are demonstrated through a comprehensive evaluation using various metrics.

References

Agrawal, A., Fu, W., & Menzies, T. (2018). What is wrong with topic modeling? And how to fix it using search-based software engineering. Information and Software Technology, 98, 74–88. https://doi.org/10.1016/j.infsof.2018.02.005

Albalawi, R., Yeap, T. H., & Benyoucef, M. (2020). Using Topic Modeling Methods for Short-Text Data: A Comparative Analysis. Frontiers in Artificial Intelligence, 3. https://doi.org/10.3389/frai.2020.00042

Alghamdi, R., & Alfalqi, K. (2015). A Survey of Topic Modeling in Text Mining. International Journal of Advanced Computer Science and Applications, 6(1). https://doi.org/10.14569/ijacsa.2015.060121

AlSumait, L., Barbará, D., Gentle, J., & Domeniconi, C. (2009). Topic Significance Ranking of LDA Generative Models. In Machine Learning and Knowledge Discovery in Databases, 67–82. Springer Berlin Heidelberg. http://dx.doi.org/10.1007/978-3-642-04180-8_22

Aouadi, H., Khemakhem, M. T., & Jemaa, M. B. (2015). An LDA Topic Model Adaptation for Context-Based Image Retrieval. In Lecture Notes in Business Information Processing, 69–80. Springer International Publishing. http://dx.doi.org/10.1007/978-3-319-27729-5_6

Arun, R., Suresh, V., Veni Madhavan, C. E., & Narasimha Murthy, M. N. (2010). On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations. In Advances in Knowledge Discovery and Data Mining, 391–402. Springer Berlin Heidelberg. http://dx.doi.org/10.1007/978-3-642-13657-3_43

Bianchi, F., Terragni, S., & Hovy, D. (2021). Pre-training is a Hot Topic: Contextualized Document Embeddings Improve Topic Coherence. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). http://dx.doi.org/10.18653/v1/2021.acl-short.96

Bide, P., Magotra, V., Jain, J., Rao, P., & Jain, K. (2021, July 8). Comprehensive Survey on Techniques of Topic Evolution Mining. 2021 6th International Conference on Communication and Electronics Systems (ICCES). http://dx.doi.org/10.1109/icces51350.2021.9488965

Blei, D. (2011, August 21). Probabilistic topic models. Proceedings of the 17th ACM SIGKDD International Conference Tutorials. http://dx.doi.org/10.1145/2107736.2107741

Blei, D. M., Jordan, M. I., & Ng, A. Y. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3(1), 993–1022.

Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. Proceedings of the 23rd International Conference on Machine Learning - ICML '06. http://dx.doi.org/10.1145/1143844.1143859

Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of Science. The Annals of Applied Statistics, 1(1). https://doi.org/10.1214/07-aoas114

Cao, J., Xia, T., Li, J., Zhang, Y., & Tang, S. (2009). A density-based method for adaptive LDA model selection. Neurocomputing, 72(7–9), 1775–1781. https://doi.org/10.1016/j.neucom.2008.06.011

Chen, F., Xie, S., Li, X., Li, S., Tang, J., & Wang, T. (2019, July). What Topics Do Images Say: A Neural Image Captioning Model with Topic Representation. 2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW). http://dx.doi.org/10.1109/icmew.2019.00083

Chen, Y., Zhang, H., Liu, R., Ye, Z., & Lin, J. (2019). Experimental explorations on short text topic mining between LDA and NMF based Schemes. Knowledge-Based Systems, 163, 1–13. https://doi.org/10.1016/j.knosys.2018.08.011

Chen, Z., & Doss, H. (2019). Inference for the Number of Topics in the Latent Dirichlet Allocation Model via Bayesian Mixture Modeling. Journal of Computational and Graphical Statistics, 28(3), 567–585. https://doi.org/10.1080/10618600.2018.1558063

Dash, S. K., Acharya, S., Pakray, P., Das, R., & Gelbukh, A. (2019). Topic-Based Image Caption Generation. Arabian Journal for Science and Engineering, 45(4), 3025–3034. https://doi.org/10.1007/s13369-019-04262-2

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41(6), 391–407. https://doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9

Dieng, A. B., Ruiz, F. J. R., & Blei, D. M. (2020). Topic Modeling in Embedding Spaces. Transactions of the Association for Computational Linguistics, 8, 439–453. https://doi.org/10.1162/tacl_a_00325

Feng, Y., & Lapata, M. (2010). Topic Models for Image Annotation and Text Illustration. Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL, 831–839. http://www.aclweb.org/anthology/N10-1125

Gan, J., & Qi, Y. (2021). Selection of the Optimal Number of Topics for LDA Topic Model—Taking Patent Policy Analysis as an Example. Entropy, 23(10), 1301. https://doi.org/10.3390/e23101301

Hasan, M., Rahman, A., Karim, Md. R., Khan, Md. S. I., & Islam, Md. J. (2020). Normalized Approach to Find Optimal Number of Topics in Latent Dirichlet Allocation (LDA). In Advances in Intelligent Systems and Computing, 341–354. Springer Singapore. http://dx.doi.org/10.1007/978-981-33-4673-4_27

Hoffman, M. D., Blei, D. M., & Bach, F. (2010). Online Learning for Latent Dirichlet Allocation. Proceedings of the 23rd International Conference on Neural Information Processing Systems, 1, 856–864. https://dl.acm.org/doi/10.5555/2997189.2997285

Hofmann, T. (1999, August). Probabilistic latent semantic indexing. Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. http://dx.doi.org/10.1145/312624.312649

Hotho, A., Nürnberger, A., & Paaß, G. (2005). A Brief Survey of Text Mining. Journal for Language Technology and Computational Linguistics, 20(1), 19–62. https://doi.org/10.21248/jlcl.20.2005.68

Kang, N. N., Wang, X. F., & Zhang, R. R. (2014). Image Classification Based on Color Topic Model. Applied Mechanics and Materials, 556–562, 4770–4773. https://doi.org/10.4028/www.scientific.net/amm.556-562.4770

Kherwa, P., & Bansal, P. (2018). Topic Modeling: A Comprehensive Review. ICST Transactions on Scalable Information Systems, 0(0), 159623. https://doi.org/10.4108/eai.13-7-2018.159623

Khurana, D., Koli, A., Khatter, K., & Singh, S. (2022). Natural language processing: state of the art, current trends and challenges. Multimedia Tools and Applications, 82(3), 3713–3744. https://doi.org/10.1007/s11042-022-13428-4

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. Nature, 401(6755), 788–791. https://doi.org/10.1038/44565

Lewis, C. M., & Grossetti, F. (2022). A Statistical Approach for Optimal Topic Model Identification. Journal of Machine Learning Research, 23(58), 1–20. https://dl.acm.org/doi/abs/10.5555/3586589.3586647

Li, W., & McCallum, A. (2006). Pachinko allocation. Proceedings of the 23rd International Conference on Machine Learning - ICML '06. http://dx.doi.org/10.1145/1143844.1143917

Liu, P., Li, L., Heng, W., & Wang, B. (2012, October). HLDA based text clustering. 2012 IEEE 2nd International Conference on Cloud Computing and Intelligence Systems. http://dx.doi.org/10.1109/ccis.2012.6664628

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. arXiv. arXiv. https://arxiv.org/abs/1301.3781

Miller, S.J., Howard, J., Adams, P., Schwan, M., and Slater, R., 2020. Multi-modal classification using images and text. SMU Data Science Review, 3(3), p.6. https://scholar.smu.edu/datasciencereview/vol3/iss3/6

Pan, Z., Liu, Y., Liu, G., Guo, M., & Li, Y. (2018b). Topic network: topic model with deep learning for image classification. Journal of Electronic Imaging, 27(03), 1. https://doi.org/10.1117/1.jei.27.3.033009

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

Peikert, S., Kubach, C., Al Qundus, J., Sandra Vu, L. D., & Paschke, A. (2021, November 29). Objective Functions to Determine the Number of Topics for Topic Modeling. The 23rd International Conference on Information Integration and Web Intelligence. http://dx.doi.org/10.1145/3487664.3487710

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global Vectors for Word Representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). http://dx.doi.org/10.3115/v1/d14-1162

Phan, X.-H., Nguyen, L.-M., & Horiguchi, S. (2008, April 21). Learning to classify short and sparse text & web with hidden topics from large-scale data collections. Proceedings of the 17th International Conference on World Wide Web. http://dx.doi.org/10.1145/1367497.1367510

Porteous, I., Newman, D., Ihler, A., Asuncion, A., Smyth, P., & Welling, M. (2008, August 24). Fast collapsed Gibbs sampling for latent Dirichlet allocation. Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. http://dx.doi.org/10.1145/1401890.1401960

Porter, M. F. (1980). An algorithm for suffix stripping. Program: Electronic Library and Information Systems, 14(3), 130–137. https://doi.org/10.1108/eb046814

Qiang, J., Qian, Z., Li, Y., Yuan, Y., & Wu, X. (2022). Short Text Topic Modeling Techniques, Applications, and Performance: A Survey. IEEE Transactions on Knowledge and Data Engineering, 34(3), 1427–1445. https://doi.org/10.1109/tkde.2020.2992485

Rehurek, R., & Sojka, P. (2011). Gensim: Python framework for vector space modelling. NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic, 3(2).

Röder, M., Both, A., & Hinneburg, A. (2015, February 2). Exploring the Space of Topic Coherence Measures. Proceedings of the Eighth ACM International Conference on Web Search and Data Mining. http://dx.doi.org/10.1145/2684822.2685324

Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (2004). The Author-Topic Model for Authors and Documents. Proceedings of the Twentieth Conference on Uncertainty in Artificial Intelligence, 487–494. https://arxiv.org/abs/1207.4169

Rüdiger, M., Antons, D., Joshi, A. M., & Salge, T.-O. (2022). Topic modeling revisited: New evidence on algorithm performance and quality metrics. PLOS ONE, 17(4), e0266325. https://doi.org/10.1371/journal.pone.0266325

Sadigov, R., Yıldırım, E., Kocaçınar, B., Patlar Akbulut, F. and Catal, C., 2023. Deep learning-based user experience evaluation in distance learning. Cluster Computing, pp.1-13. https://doi.org/10.1007/s10586-022-03918-3

Sbalchiero, S., & Eder, M. (2020). Topic modeling, long texts and the best number of topics. Some Problems and solutions. Quality & Quantity, 54(4), 1095–1108. https://doi.org/10.1007/s11135-020-00976-w

Terragni, S., Fersini, E., Galuzzi, B. G., Tropeano, P., & Candelieri, A. (2021). OCTIS: Comparing and Optimizing Topic Models is Simple! Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations. http://dx.doi.org/10.18653/v1/2021.eacl-demos.31

Tsapatsoulis, N., 2022, September. Classification of Instagram photos: topic modeling vs transfer learning. In Proceedings of the 12th Hellenic Conference on Artificial Intelligence (pp. 1-7). https://doi.org/10.1145/3549737.3549759

Wang, C., Richardson, Y., & Sander, R. (2019). Unsupervised Image Clustering and Topic Modeling for Accelerated Annotation. ResearchGate, 2019. https://doi.org/10.13140/RG.2.2.27176.52484

Weisser, C., Gerloff, C., Thielmann, A., Python, A., Reuter, A., Kneib, T. and Säfken, B., 2023. Pseudo-document simulation for comparing LDA, GSDMM, and GPM topic models on short and sparse text using Twitter data. Computational Statistics, 38(2), pp.647-674. https://doi.org/10.1007/s00180-022-01246-z

Wu, X., Dong, X., Nguyen, T.T. and Luu, A.T., 2023, July. Effective neural topic modeling with embedding clustering regularization. In International Conference on Machine Learning (pp. 37335-37357). PMLR. https://doi.org/10.48550/arXiv.2306.04217

Xia, W., & Doss, H. (2020). Scalable Hyperparameter Selection for Latent Dirichlet Allocation. Journal of Computational and Graphical Statistics, 29(4), 875–895. https://doi.org/10.1080/10618600.2020.1741378

Zhu, Z., Xue, Z., & Yuan, Z. (2018, October). Topic-Guided Attention for Image Captioning. 2018 25th IEEE International Conference on Image Processing (ICIP). http://dx.doi.org/10.1109/icip.2018.8451083

Zia, U., Riaz, M. M., Ghafoor, A., & Ali, S. S. (2019). Topic sensitive image descriptions. Neural Computing and Applications, 32(14), 10471–10479. https://doi.org/10.1007/s00521-019-04587-x

Appendix

Table 1: Analysis of Recent Literature

S.No.	Title	Author, Year, and Journal	Methodology	Research Contribution
1.	Effective Neural Topic Modeling with Embedding Clustering Regularization. (Wu et al., 2023)	Wu, X., Dong, X., Nguyen, T.T. and Luu, A.T., 2023, July. Effective neural topic modeling with embedding clustering regularization. In International Conference on Machine Learning (pp. 37335-37357). PMLR.	The authors introduce ECRTM, a novel neural topic model to mitigate "topic collapsing" problems. They present Embedding Clustering Regularization (ECR) as a solution to prevent the convergence of topic embeddings and to maintain distinct word semantics within topics. ECRTM, enriched by the inclusion of ECR, demonstrates the ability to generate diverse and coherent topics and high-quality document-topic distributions. The authors' contributions encompass the introduction of ECR, its seamless integration into ECRTM, and a series of experiments showcasing its effectiveness in addressing the topics of collapse.	The study evaluates topic models regarding topic quality and doc-topic distribution quality. Topic quality is assessed using Topic Coherence (CV) and Topic Diversity (TD) metrics, with a Wikipedia article collection as the reference corpus for fair comparisons. Document clustering quality is evaluated with Purity and NMI metrics. Models compared include LDA, KM, DVAE, WLDA, ETM, HyperMiner, NSTM, and WeTe.
2.	Pseudo-document simulation for comparing LDA, GSDMM, and GPM topic models on short and sparse text using Twitter data. (Weisser et al., 2023)	Weisser, C., Gerloff, C., Thielmann, A., Python, A., Reuter, A., Kneib, T. and Säfken, B., 2023. Pseudo-document simulation for comparing LDA, GSDMM, and GPM topic models on short and sparse text using Twitter data. Computational Statistics, 38(2), pp.647-674.	The study explores the challenges posed by short and sparse social media texts, such as Twitter, for the conventional Latent Dirichlet Allocation (LDA) topic model. It compares LDA with two models designed for sparse data: the Gibbs Sampler Dirichlet Multinomial Model (GSDMM) and the Gamma Poisson Mixture Model (GPM). A novel evaluation method involving pseudo-document simulation is introduced. In a case study involving Covid-19-related tweets, the study finds that standard coherence scores, commonly used for topic model evaluation, are ineffective as evaluation metrics.	Coherence scores indicate LDA outperforms GPM and GSDMM despite the latter being designed for sparse text. However, using Pseudo-Document Simulation, GSDMM and GPM outperform LDA in model comparison, showcasing the effectiveness of this method for researchers to compare topic models.
3.	Deep learning-based user experience evaluation in distance learning. (Sadigov et al.,	Sadigov, R., Yıldırım, E., Kocaçınar, B., Patlar Akbulut, F. and Catal, C., 2023. Deep learning-based user	This study assesses the effectiveness of e-learning during the Covid-19 pandemic using over 160,000 tweets related to changes in education. Based on the LDA algorithm, sentiment analysis and topic modeling were employed. An	Various text preprocessing techniques were applied for content cleaning, including tokenization, normalization, stop-word removal, and stemming.

	2023)	experience evaluation in distance learning. Cluster Computing, pp.1-13.	extended short-term memory-based sentiment analysis model with word2vec embedding gauged Twitter users' opinions on distance education. An LDA-based topic model was also constructed to identify discussed topics on Twitter.	Word embedding techniques, particularly Word2Vec, were used to construct neural language models based on word vectors. Deep learning-based sentiment models and topic modeling with the LDA algorithm were then developed. Notably, word embedding enhanced the performance of the LSTM-based sentiment model. The LSTM model using Word2Vec achieved the highest validation accuracy at 75.9%.
4.	Classification of Instagram photos: topic modeling vs transfer learning. (Tsapatsoulis et al., 2022)	Tsapatsoulis, N., 2022, September. Classification of Instagram photos: topic modeling vs transfer learning. In Proceedings of the 12th Hellenic Conference on Artificial Intelligence (pp. 1-7).	The authors investigate using pre-trained deep learning models, like Resnet-50, for transfer learning, including image retrieval. Their prior work introduced topic modeling for Instagram photo retrieval via hashtags. This study compares content-based image classification (transfer learning) with hashtag-based topic modeling across 24 concepts.	On average, topic modeling-based classification outperforms transfer learning-based classification in top-1 and top-5 accuracy (83.5% vs 71.7% for top-1 and 97.5% vs 87.4% for top-5). This result indicates that text-based image retrieval still performs better than content-based image retrieval, even in the deep learning era.
5.	Topic modeling revisited: New evidence on algorithm performance and quality metrics. (Rüdiger, et al., 2022)	Rüdiger, M., Antons, D., Joshi, A.M., Salge, T.-O., 2022. PLOS ONE 17, e0266325.	In their study, the authors conduct a two-phase evaluation: a comprehensive evaluation of the topic modeling algorithms to determine the most suitable performant model and a comprehensive evaluation of the quality metrics to determine the interplay of performance, hyperparameter selection, and metric results. The aim is to identify candidate algorithms and metrics yielding the best clustering performance in terms of accuracy. The former is achieved through a holistic comparison of all commonly used, non-application-specific topic modeling algorithms and an in-depth assessment of their	Evaluation over a curated dataset of English Wikipedia articles as documents and corresponding category information as ground-truth labels yield the Gibbs-Sampling LDA method to be dominant for large values of K over 20 and NMF for the small values of K in terms of both clustering accuracy and result scattering, consistent with results observed in existing studies. In terms of the performance trend, it was observed that an

			relative performance. The comparison is made against a known clustering and thus enables an unbiased evaluation of results. For the latter, the authors analyze the relationship between existing metrics and the known clustering and thus objectively determine under what conditions these algorithms may be utilized effectively.	increase in the value of K negatively impacts the performance. Further, internal quality metrics are unreliable in determining optimal hyperparameters based on evaluating the correlation with the target external cluster validity measures.
6.	Short Text Topic Modeling Techniques, Applications, and Performance: A Survey. (Qiang, et al., 2022)	Qiang, J., Qian, Z., Li, Y., Yuan, Y., Wu, X., 2022. IEEE Transactions on Knowledge and Data Engineering 34, 1427–1445.	Discussed the application of topic modeling over short-text data, providing a comprehensive review of short-text topic modeling techniques. Discusses three categories of techniques: Dirichlet Multinomial Mixture (DMM), Global word co-occurrences, and self-aggregation. Evaluate state-of-the-art methods over real-world datasets and compare them with standard topic models using classification, clustering, and coherence metrics.	The paper provides insight into how to use topic modeling algorithms over short text data. This survey paper has analyzed various techniques. Comparison results showed that the different techniques resulted better for different data sets. The authors developed the first comprehensive open-source library, STTM, for use in Java that integrates all surveyed algorithms in one interface along with benchmark datasets.
7.	A Statistical Approach for Optimal Topic Model Identification. (Lewis & Grossetti, 2022)	Lewis, C. M., & Grossetti, F. (2022). Journal of Machine Learning Research, 23(58), 1–20	Proposed a method for determining the optimal LDA configuration using parametric tests based on the multinomial distribution of the original LDA. Based on the conjecture, the minimal number of topics defining a corpus (K) should define the corpus better than any other specification. Chi-square tests followed by topic stability determination are used to identify the optimal number of topics. Two more tests are conducted to address overall model adequacy using a series of chi-square tests. Segregates topics into informative and uninformative categories.	The proposed methodology was evaluated on the U.S. Presidential Inaugural Address Corpus, and values for K were consistent with observations. The results were validated through a simulation study and compared to standard heuristic metrics like the perplexity index. The proposed method outperformed when compared with the perplexity index, and uninformative topics were identified.
8.	Objective Functions to Determine the Number of Topics for Topic Modeling. (Peikert, et al.,	Peikert, S., Kubach, C., Al Qundus, J., Sandra Vu, L.D., Paschke, A., 2021., The 23rd	Proposed some objective functions for evaluating topic models to determine optimal hyperparameters. The objective functions used in the approach improvise over the topic	Empirical studies conducted with the NMF algorithm using the proposed objective functions over different datasets yield promising

	2021)	International Conference on Information Integration and Web Intelligence. ACM, New York, NY, USA.	coherence metrics proposed by Röder et al., 2015, which are used for internal topic evaluation. These are combined with external topic evaluation functions, which evaluate topic similarity across pairs of generated topics. The combination allows for holistic evaluation with an extended focus on interpretability.	results for determining the optimal values of the number of topics.
9.	Selection of the Optimal Number of Topics for LDA Topic Model—Taking Patent Policy Analysis as an Example. (Gan & Qi, 2021)	Gan, J., Qi, Y., 2021. Selection of the Optimal Number of Topics for LDA Topic Model—Taking Patent Policy Analysis as an Example. Entropy 23, 1301.	A general evaluation metric was developed to improve the topic model evaluation. Their comprehensive metric focuses on judging the quality based on stability, predictive ability, and topic isolation of topic models. It combines the information given by perplexity, isolation (based on Kullback-Leibler or Jensen-Shannon divergence), stability (based on ranking by average Jaccard index), and coincidence (based on coherence measurements) into a topic score to determine the best value. The authors propose a methodology to determine the most suitable value for the number of topics for LDA topic models by calculating the degree of model stability across different topic numbers. The methodology involves generating an initial topic model based on the complete dataset, randomly selecting samples from the dataset to generate multiple topic models, and calculating the stability and coincidence of each model with the test model. The remaining samples are used as the training set, and the isolation and perplexity for each test sample are computed. Topic superiority scores of sample models are then averaged across varying values of k, and the minimum value of k whose sample models have a significant average topic superiority score is selected as the optimal value.	The authors evaluate their proposal over standard datasets such as 20 Newsgroups, WOS 46985, and AG News and evaluate multiple variants of LDA based on online variational inference, collapsed Gibbs-sampling inference, and an LDA model based on GloVe word-embeddings and other topic models such as HDP, and verify the efficacy of the method regarding manual cluster inspection by projecting the clustering to a low dimensional space using t-SNE. The choice of the number of topics by the proposed method yields high-quality clustering with low fuzzy boundaries when further examined against known clustering assignments.
10.	Normalized Approach to Find Optimal Number of Topics in Latent Dirichlet Allocation	Hasan, Mahedi, Anichur Rahman, Md. Razaul Karim, Md. Saikat Islam Khan and Md.	The research focused on evaluating the extent of the influence of perplexity and coherence over the quality of the topics produced by LDA models to improve the F-	The normalized coherence score metric outperforms other methods regarding the topics' coherence and

	(LDA). (Hasan, et al., 2020)	Jahidul Islam. 2020. Advances in Intelligent Systems and Computing. Springer Singapore, Singapore, 341–354.	Measure for examining topic quality in a topic classification task. The normalized coherence score is calculated by dividing the coherence score of a particular number of topics by the maximum coherence score across all tested numbers of topics. This normalization step helps to reduce the influence of the size of the corpus on the coherence score and makes it easier to compare models trained on different datasets. They compare their approach to several other methods for selecting the number of topics, including the perplexity measure, the topic coherence measure, and the topic uniqueness measure for different datasets. They found that the best topics were determined based on values of the normalized absolute perplexity (NAP) and normalized absolute coherence (NAC) (UMass) metrics.	interpretability.
11.	Scalable Hyperparameter Selection for Latent Dirichlet Allocation. (Xia & Doss, 2020)	Xia, W., Doss, H., 2020. Scalable Hyperparameter Selection for Latent Dirichlet Allocation. Journal of Computational and Graphical Statistics 29, 875–895	The paper proposes a novel method for selecting LDA hyperparameters based on a hierarchical Bayesian framework. The method involves setting hyperpriors on the hyperparameters and using variational inference to estimate the posterior distribution of the hyperparameters. The proposed method is scalable, and experiments conducted by the authors demonstrate its ability to handle large datasets using stochastic optimization techniques.	The authors evaluate the proposed method with several existing methods for hyperparameter selection in LDA. The evaluation shows that the proposed method outperforms the existing methods regarding topic coherence and topic diversity. The proposed method is applied to a real-world dataset of scientific publications to extract topics and analyze research trends. The results show that the proposed method can identify meaningful topics and provide insights into research trends in the field.
12.	Using Topic Modeling Methods for Short-Text Data: A Comparative Analysis. (Albalawi, et al., 2020)	Albalawi, R., Yeap, T.H., Benyoucef, M., 2020. Frontiers in Artificial Intelligence 3.	Discussed approaches to discover topics in short-text data. Compared five topic models and algorithms: LSA, LDA, NMF, Random Projection (RP), and Principal Component Analysis (PCA) over two datasets. An evaluation was conducted using methods based on the topic quality	Over statistical metrics, promising results for LDA and NMF, with more meaningful extracted topics.

			and statistical metrics like recall, precision, F-score, and topic coherence.	
13.	Topic modeling, long texts, and the best number of topics. Some problems and solutions. (Sbalchiero & Eder, 2020)	Stefano Sbalchiero, Maciej Eder, 2020. Quality & Quantity	Presents the results of different experiments on the problem of the model fitting process in topic modeling and its accuracy on long text datasets. The research was conducted via three experiments applied to text chunks of different lengths. The experiment was performed on a large corpus of English novels, then on the reduced corpus of 10 novels, and lastly on an equal number of text chunks. The author calculated the log-likelihood for each sample size to identify the best number of topics.	The results were stable up to some text length, and then inferred optimal topics decreased when the size of the text chunk increased, showing a clear relationship between the optimal number of topics 'K' and the size of the sample text presenting the Sbalchiero-Eder rule. According to this, the best number of topics is inversely proportional to the length of the text chunks, i.e., the larger the portions of the text, the lower the best number of topics. The equation presents this: $y=f(x)$, where y is the best number of topics and x is the size in word tokens of the text chunks. The function f is a steep decreasing function that may be a power or logarithmic model.
14.	Multi-Modal Classification Using Images and Text. (Miller et al., 2020)	Miller, S.J., Howard, J., Adams, P., Schwan, M., and Slater, R., 2020. Multi-modal classification using images and text. SMU Data Science Review, 3(3), p.6.	This paper suggests enhancing image classification accuracy by integrating natural language understanding with image metadata. Image features are typically used for classification, but this approach incorporates descriptions, titles, and tags through a multi-modal model. The novelty lies in utilizing natural language understanding and transfer learning to learn from external image-related features.	The research results indicated that when ResNet-50 image feature extraction was combined with Universal Sentence Encoder embeddings, the Top 5 error rate was 73.05%, and the Top 1 error rate was 54.65%. This combination showed a 1.56% enhancement compared to the benchmark results.
15.	Topic sensitive image descriptions. (Zia et al., 2019)	Zia, U., Riaz, M.M., Ghafoor, A., Ali, S.S., 2019. 32, 10471–10479. Neural Computing and Applications.	Proposal of a framework for generating topic-sensitive descriptions for images. Incorporates topic-sensitive word embeddings into the captioning process, which combines associated topics with word embeddings. Uses HDP for topic identification and R-CNN for image feature extraction.	The proposed model reports significant improvement in most evaluation metrics except for BLEU-4 compared to existing state-of-the-art models on Flickr8K, Flickr30K, and MSCOCO datasets. The proposed model is compared further to

				measure the performance of the generated topic-sensitive description of images with state-of-the-art models. The result shows that the model can generate sensible and more relevant image descriptions.
16.	Topic Modeling: A Comprehensive Review. (Kherwa & Bansal, 2019)	Kherwa, P., & Bansal, P. (2018). ICST Transactions on Scalable Information Systems, 159623.	It has provided a comprehensive review of topic modeling and its applications. They discussed several versions of topic models, including classification hierarchy: Latent Semantic Analysis (LSA), Latent Dirichlet Allocation (LDA), Probabilistic Latent Semantic Analysis (PLSA), and Non-negative Matrix Factorization (NMF). Apart from topic models, they also discussed posterior inference techniques, LDA's evolution models, and their applications in different technology areas. It also elaborates on the challenges of topic modeling in detail.	The paper provides a good insight into various topic modeling algorithms and discusses various applications.
17.	Unsupervised Image Clustering and Topic Modeling for Accelerated Annotation. (Wang, et al., 2019)	Wang, C., Richardson, Y., Sander, R., 2019. ResearchGate 2019.	Proposed an unsupervised approach for image clustering using latent features from image topic models. Image features obtained via SIFT or CNNs, followed by dimensionality reduction via K-Means to generate bag-of-words representation in lower dimensional space. Bag-of-visual-words representation utilized by topic models to obtain latent topics used for organization.	Using latent topic embeddings, the proposed method can group topics of similar ground truth labels. The model is highly dependent on feature embeddings, which can hamper the performance if not appropriately considered.
18.	Topic-Based Image Caption Generation. (Dash et al., 2019)	Dash, S.K., Acharya, S., Pakray, P., Das, R., Gelbukh, A., 2019. Arabian Journal for Science and Engineering 45, 3025–3034.	Proposed a method for image captioning by including image topics alongside object attributes. Followed the approach of combining multimodal features in an encoder-decoder framework.	The proposed method outperforms several state-of-the-art image captioning models on various evaluation metrics, including BLEU, METEOR, ROGUE-L, and CIDEr. In addition, the authors conduct a human evaluation, showing that their proposed method generates more informative and diverse captions than other methods and has potential applications in various domains,

				including image and video captioning and visual question answering.
19.	Experimental explorations on short text topic mining between LDA and NMF based Schemes. (Chen, et al., 2019)	Chen, Y., Zhang, H., Liu, R., Ye, Z., Lin, J., 2019. Knowledge-Based Systems 163, 1–13.	The experiments conducted by the authors are divided into two parts. In the first part, the authors compared the basic LDA and NMF models with different experimental settings on several public short text datasets. The datasets included news headlines, tweets, and product reviews. They evaluated the models based on their ability to generate meaningful topics and their coherence using metrics such as perplexity, coherence score, and topic diversity. In the second part, the authors proposed a novel KGNMF model, which leverages external knowledge as a semantic regulator with low-rank formalizations. They conducted extensive experiments on three representative corpora with currently typical short text topic models to demonstrate the effectiveness of KGNMF.	NMF performs better than LDA in basic settings on the news headlines and tweets datasets. LDA performed better than NMF on the product reviews dataset. The authors attributed this to the sparsity of the product reviews dataset, which may have resulted in better performance for NMF.
20.	Inference for the Number of Topics in the Latent Dirichlet Allocation Model via Bayesian Mixture Modeling. (Chen & Doss, 2019)	Chen, Zhe, and Hani Doss. 2019 Journal of Computational and Graphical Statistics 28 (3): 567–585.	The Optimal number of topics is determined by following the Bayesian approach of imposing a prior on the number of topics and estimating the mixture distribution on latent variables of the model. Estimation is achieved by a proposed variant of the pseudo-marginal Metropolis-Hastings algorithm. The topic posterior depended on a fixed number of topics computed to estimate the general topic posterior in the assumed model with a prior distribution over the number of topics.	The model outperformed the other methods in accuracy and robustness. It is concluded that the proposed model could accurately estimate the number of topics, even in cases where the data was highly imbalanced or the correct number of topics was unknown.
21.	Topic-Guided Attention for Image Captioning. (Zhu, et al., 2018)	Zhu, Z., Xue, Z., Yuan, Z., 2018., 2018 25th IEEE International Conference on Image Processing (ICIP).	Proposed an attention mechanism for automatic image captioning by integration of image topics. Integrates image topics into the attention model in an encoder-decoder model to guide selecting relevant image features.	The proposed method outperforms several state-of-the-art image captioning models on benchmark datasets and has potential applications in various domains, including image and video captioning and visual question answering.

22.	What is wrong with topic modeling? And how to fix it using search-based software engineering. (Agrawal, et al., 2018)	Agrawal, A., Fu, W., Menzies, T., 2018. Information and Software Technology 98, 74–88.	Proposed a method to generate more stable LDA distributions and how it can be used for further analysis. They used Differential Evolution (DE) to tune the LDA's parameters. They evaluated the proposed model on different datasets, including Stackoverflow, SE papers, and software defect reports from NASA. Different implementations of LDA were carried out for different kinds of LDAs (VEM, Gibbs Sampling) and evaluated via topic stability and text mining classification accuracy.	The result shows that the standard LDA exhibits considerable topic instability. With LDADE's tunings, cluster instability can be reduced, leading to improved performances for supervised and unsupervised learning.
23.	Topic network: topic model with deep learning for image classification. (Pan, et al., 2018)	Pan, Z., Liu, Y., Liu, G., Guo, M., Li, Y., 2018. Journal of Electronic Imaging.	It proposed a deep learning network that combines topic models with CNNs for image classification. Replaces pooling layers with topic models for dimensionality reduction without significant loss of information.	Both the topic models LDA and MTRF performed well in image classification with original size. Compared to other approaches, it's been shown that any topic model can be easily applied to the topic layer, making the proposed framework flexible and extensible.

Developing an Efficient Topic Modeling Framework for Image Classification

by Astha Goyal

Submission date: 22-Nov-2023 07:04AM (UTC+0530)
Submission ID: 2235690730
File name: Efficient_Topic_Modeling_Framework_for_Image_Classification.docx (230.19K)
Word count: 11763
Character count: 69967

SHORT SYNOPSIS
by

Ms. Astha Goyal
Registration number: 21/Ph.D/046

Under the supervision of

Supervisor
Dr. Indu Kashyap
Professor, CSE

SCHOOL OF ENGINEERING & TECHNOLOGY
MANAV RACHNA INTERNATIONAL INSTITUTE OF RESEARCH & STUDIES
SECTOR-43, SURAJKUND –DELHI ROAD, FARIDABAD – 121001

ABSTRACT

Topic modeling is a powerful tool with many applications in fields such as text summarization, sentiment analysis, and reviewing of vast literature. In such contexts, topic models are used to discern and extract the relevant topics or themes within a given corpus of text, enabling the creation of summaries, the analysis of sentiment, or the synthesis of research findings. However, topic models are not limited to application over textual data. They can also be employed over other forms of data, such as multimedia content, for tasks such as classification and captioning.

This research investigates the effectiveness of utilizing topic models on image captions to classify images. Existing approaches for image classification focus solely on visual information and fail to use information from other modalities, as well as abstract information

MANAV RACHNA
Vidyanatariksha

Dr. O.P. Bhalla Central Library
Manav Rachna International Institute of Research and Studies
(A deemed-to-be-university under section 3 of UGC Act 1956)

SIMILARITY VERIFICATION CERTIFICATE

Title: Developing an Efficient Topic Modeling Framework for Image Classification

Researcher: Astha Goyal
Registration Number: 21/Ph.D/046

Supervisor: Dr. Indu Kashyap
Department: CSE

This is to certify that the above mentioned Short Synopsis/Thesis was scanned for similarity index through Turnitin Software on 22/11/23 with submission ID 2235690739 and the similarity index was found to be 8 %. The complete report running into 44 Pages, 11763 Word count with File Size 5857 KB is being forwarded to the Research Supervisor/DRC/DRB for review.

Date: 23/11/2023

Rajesh Kumar 23/11/23
Dr. Rajesh Kumar
University Librarian
Dr. O.P. Bhalla Central Library
Manav Rachna International
Institute of Research & Studies
Faridabad

The complete report of the above Short Synopsis/Thesis has been reviewed and the similarity index is below the accepted norms as per the University Grants Commission (Promotion of Academic Integrity and Prevention of Plagiarism in Higher Educational Institutions) Regulations 2018. The Short Synopsis/Thesis may be considered for further processing

Research Scholar

Research Supervisor 23/11

Developing an Efficient Topic Modeling Framework for Image Classification

ORIGINALITY REPORT

Dr. Rajesh Kumar
University Librarian
Central Library
Manav Rachna International
Studies
Faridabad

8%	6%	8%	1%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	link.springer.com Internet Source	1%
2	www.mdpi.com Internet Source	1%
3	www.ncbi.nlm.nih.gov Internet Source	1%
4	www.researchgate.net Internet Source	<1%
5	Jipeng Qiang, Zhenyu Qian, Yun Li, Yunhao Yuan, Xindong Wu. "Short Text Topic Modeling Techniques, Applications, and Performance: A Survey", IEEE Transactions on Knowledge and Data Engineering, 2022 Publication	<1%
6	Stefano Sbalchiero, Maciej Eder. "Topic modeling, long texts and the best number of topics. Some Problems and solutions", Quality & Quantity, 2020 Publication	<1%

7	Astha Goyal, Indu Kashyap. "Latent Dirichlet Allocation - An approach for topic discovery", 2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON), 2022 Publication	<1 %
8	shodhgangotri.inflibnet.ac.in Internet Source	<1 %
9	Chikhaoui, Belkacem, Shengrui Wang, and Hélène Pigot. "ADR-SPLDA: Activity discovery and recognition by combining sequential patterns and latent Dirichlet allocation", Pervasive and Mobile Computing, 2012. Publication	<1 %
10	dokumen.pub Internet Source	<1 %
11	"Proceedings of International Conference on Trends in Computational and Cognitive Engineering", Springer Science and Business Media LLC, 2021 Publication	<1 %
12	Belal Abdullah Hezam Murshed, Suresha Mallappa, Jemal Abawajy, Mufeed Ahmed Naji Saif et al. "Short text topic modelling approaches in the context of big data: taxonomy, survey, and analysis", Artificial Intelligence Review, 2022 Publication	<1 %

13	Simon Rogers. "Infinite factorization of multiple non-parametric views", Machine Learning, 11/13/2009 Publication	<1 %
14	academica-e.unavarra.es Internet Source	<1 %
15	Sandeep Kumar Dash, Shantanu Acharya, Partha Pakray, Ranjita Das, Alexander Gelbukh. "Topic-Based Image Caption Generation", Arabian Journal for Science and Engineering, 2019 Publication	<1 %
16	repository.unika.ac.id Internet Source	<1 %
17	webthesis.biblio.polito.it Internet Source	<1 %
18	"Artificial Neural Networks and Machine Learning – ICANN 2019: Text and Time Series", Springer Science and Business Media LLC, 2019 Publication	<1 %
19	www.spiedigitallibrary.org Internet Source	<1 %
20	"Hybrid Artificial Intelligent Systems", Springer Science and Business Media LLC, 2023	<1 %

21	Silvio Peikert, Clemens Kubach, Jamal Al Qundus, Le Duyen Sandra Vu, Adrian Paschke. "Objective Functions to Determine the Number of Topics for Topic Modeling", The 23rd International Conference on Information Integration and Web Intelligence, 2021 Publication	<1 %
22	theses.gla.ac.uk Internet Source	<1 %
23	Wei Xia, Hani Doss. "Scalable Hyperparameter Selection for Latent Dirichlet Allocation", Journal of Computational and Graphical Statistics, 2020 Publication	<1 %
24	curve.carleton.ca Internet Source	<1 %
25	"Artificial Intelligence in Medicine", Springer Science and Business Media LLC, 2020 Publication	<1 %
26	Ma, Yinglong, Yao Wang, and Beihong Jin. "A three-phase approach to document clustering based on topic significance degree", Expert Systems with Applications, 2014. Publication	<1 %
27	www.frontiersin.org	

		<1 %
28	www.biorxiv.org Internet Source	<1 %

Exclude quotes	On	Exclude matches	< 14 words
Exclude bibliography	On		